



El uso de machine learning para la investigación en asuntos públicos

Bastián González-Bustamante

University of Oxford

Universidad de Santiago de Chile

✉ bastian.gonzalezbustamante@politics.ox.ac.uk

Presentación preparada para XII Congreso Chileno de Administración Pública
Universidad de Valparaíso, 3 al 5 de noviembre de 2021

Tabla de contenidos

1. Inteligencia artificial y machine learning
2. Reconocimiento óptico de información del servicio civil
3. NLP para eventos en América Latina
4. Training Data Lab



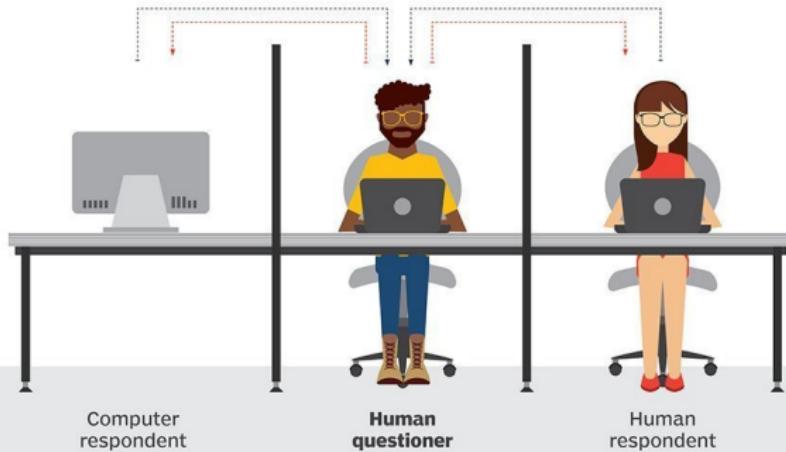
Inteligencia artificial y machine learning

Turing test

During the Turing test, the human questioner asks a series of questions to both respondents.

After the specified time, the questioner tries to decide which terminal is operated by the human respondent and which terminal is operated by the computer.

■ QUESTION TO RESPONDENTS ■ ANSWERS TO QUESTIONER



Se asocia la IA con la capacidad de crear programas que puedan realizar operaciones como los humanos, por ejemplo, razonamiento lógico y aprendizaje.

Para pasar el test de Turing sería necesario:

- Procesamiento de lenguaje natural (NLP)
- Almacenar información
- Razonamiento (econometría clásica) y aprendizaje automático (ML)
- Percepción del entorno y robótica (interactuar)

Aprendizaje automático

ML se puede considerar un subcampo de computer science que construye **algoritmos** para resolver un problema práctico básicamente usando reco-pilación de **big data** y la construcción de modelos a partir de esos datos.

- Requiere de cierta capacidad de procesamiento
- No necesariamente hay teoría para entender relaciones
- La inteligencia artificial puede aprender de relaciones subyacentes en big data

Aprendizaje supervisado para encontrar $Y = f(X)$: $(X_i, Y_i)_{i=1}^n$

Aprendizaje no supervisado (patrones): $(X_i)_{i=1}^n$

Procesamiento automático de información

Aunque actualmente existen técnicas para producir de forma automática datos sobre eventos en ciencia política y administración pública, mayoritariamente se continúa utilizando codificación manual.

La **codificación manual es costosa** en tiempo y recursos e impide garantizar la **reproducibilidad** del proceso ([Radford, 2019](#)).

Desde mediados de la década de 1990 los esfuerzos de codificación automatizada para elaborar conjuntos de datos de eventos y fenómenos de interés público han crecido. La mayoría de estos conjuntos de datos han sido construidos con fuentes abiertas de noticias y se han procesado con diccionarios pre-elaborados.

Reconocimiento óptico de información del servicio civil

Algoritmo OCR para servicio civil chileno

 Bastián González-Bustamante, Matías Astete y Berenice Orvenes

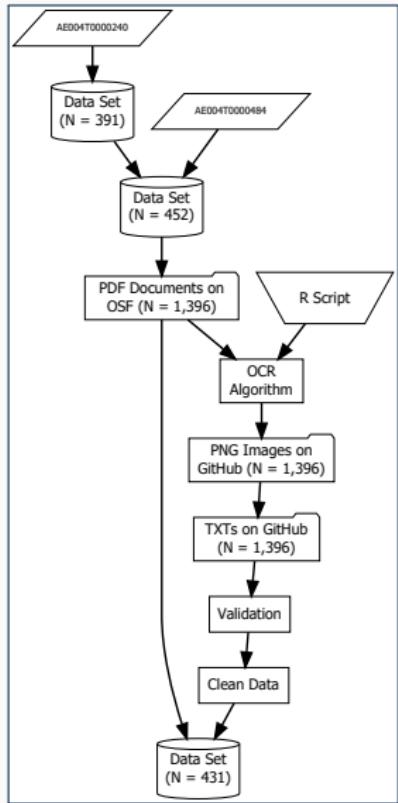
 DOI: [10.17605/OSF.IO/WBF6M](https://doi.org/10.17605/OSF.IO/WBF6M)

 training-datalab.com/projects/chilean-civil-service

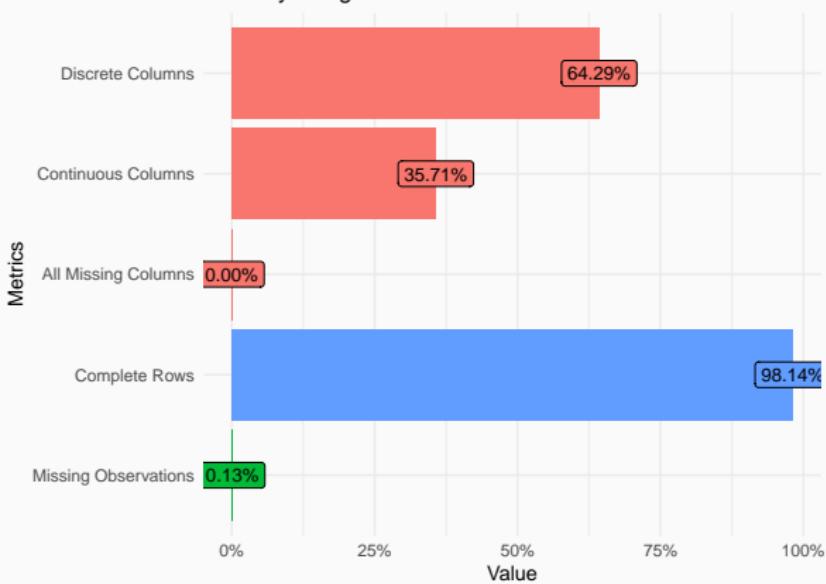
A Novel Dataset on Members of the Chilean Civil Service. Conjunto de datos con información detallada de 431 altos directivos públicos del primer nivel jerárquico del servicio civil chileno (2009-2017). Fue creado con dos solicitudes de acceso a información pública realizadas a la DNSC y una revisión de 1.396 documentos públicos, principalmente decretos y noticias institucionales. Estos documentos fueron revisados con algoritmos de minería de datos.

Artículo publicado en *Revista de Gestión Pública* (González-Bustamante et al., 2020),  código en  en  GitHub (bajo embargo hasta mediados de 2022). Preprint en inglés disponible en *SocArXiv* (González-Bustamante et al., 2021).

Proceso de recolección de información



Memory Usage: 243 Kb



Creación del conjunto de datos

Tesseract. Motor para reconocimiento óptico que se comenzó a desarrollar en 1995 en Bristol y desde 2005 está disponible como un código abierto y actualmente es usado por [Google](#) (Smith, 2007).

La aplicación de OCR sigue varias etapas. Primero, se identifican componentes conectados y se anidan los contornos y las líneas de texto. Luego, las líneas se dividen en palabras considerando espacios. Cada palabra se intenta reconocer con diccionarios de datos entrenados (*baseline*).

Este proceso implica machine learning, por tanto, se realiza una iteración con el fin de reconocer palabras que en una primera instancia no fueron identificadas. Finalmente, se revisan los espaciados difusos.

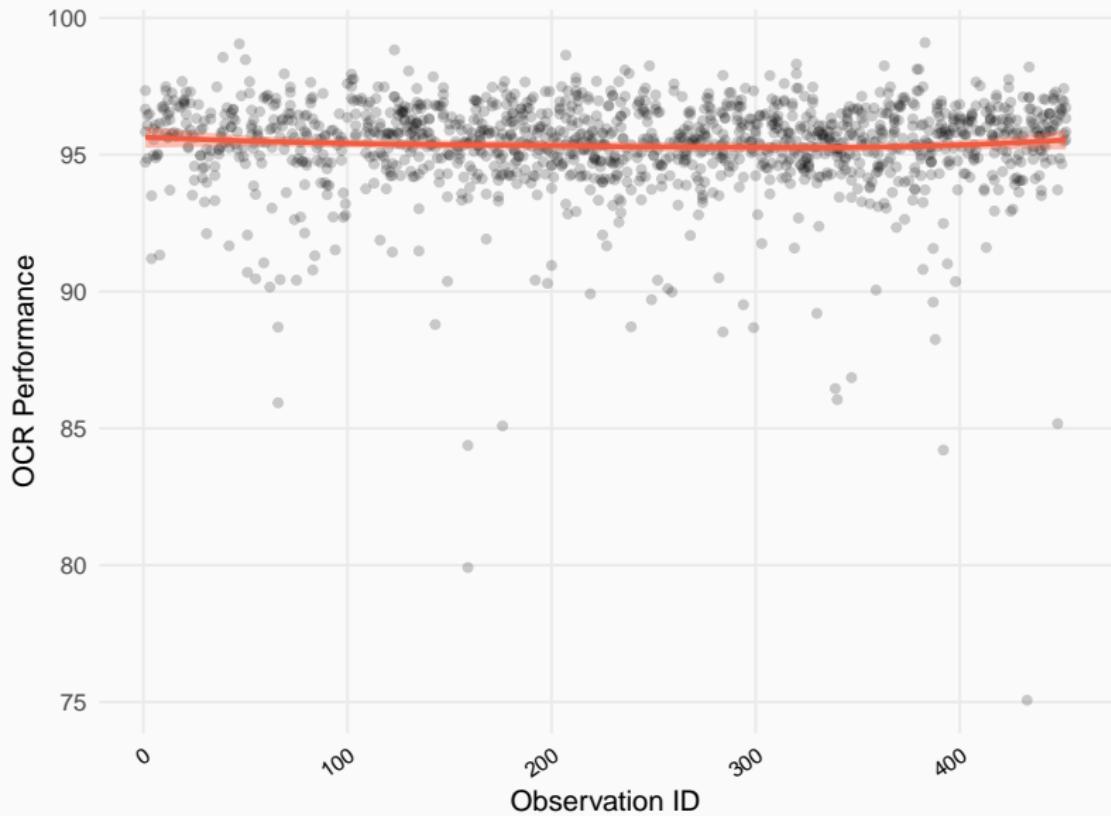
Creación, validación y anonimización

Precisión del algoritmo OCR. Se evaluó la proporción de texto que logró identificar correctamente. Se contrastan las palabras identificadas con diccionarios del idioma usados para entrenar los modelos Long Short Term Memory (LSTM) usados por Tesseract.

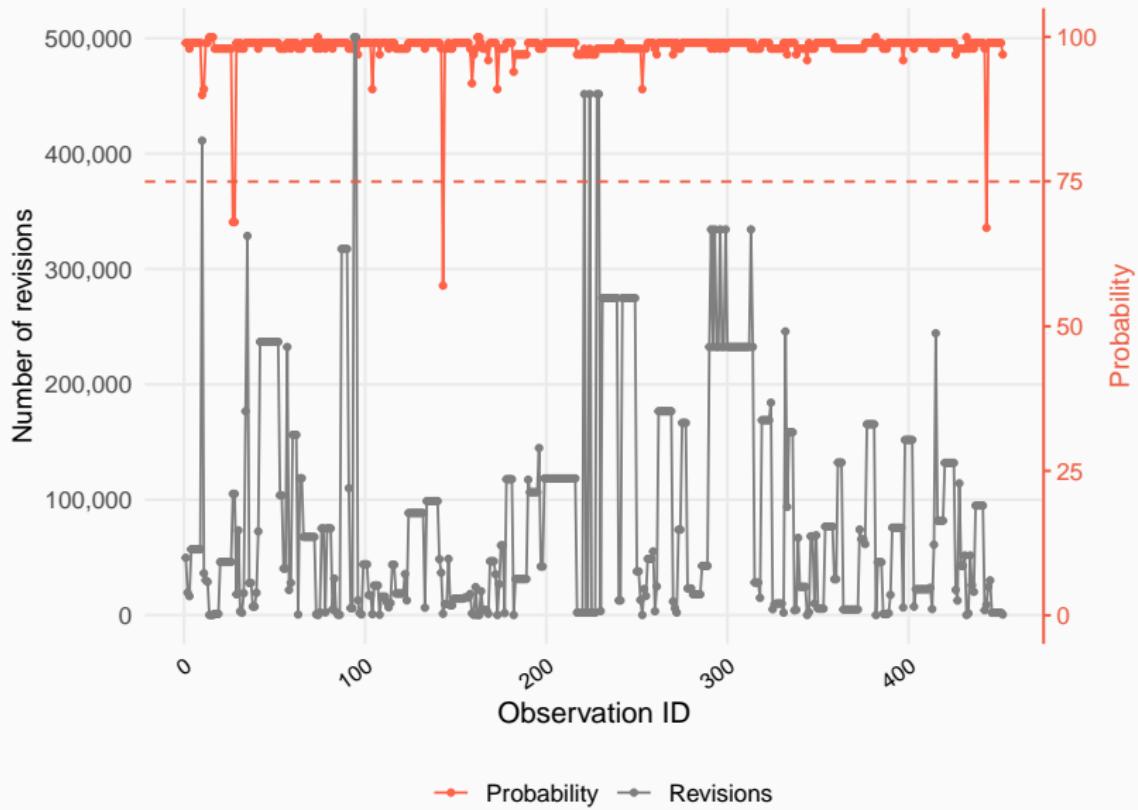
Validación automática de sexo. Usando el primer nombre de cada caso y una base de datos de nombres de diversos países del mundo, sexo y su predicción estimada creada por Casper Strömgren en 2013 ([Wais, 2016](#)). Esta base crece diariamente con datos extraídos de perfiles de redes sociales y en el momento de la revisión contaba con 111.541.298 observaciones, de las cuales 210.959 (0,19 %) correspondían a casos de Chile.

Algoritmo criptográfico. Las variables que contienen información personal fueron anonimizadas en la versión final del conjunto de datos con Secure Hash Algorithm de 256 caracteres (SHA256) basado en una función *hash* que bloquea la ingeniería-reversa ([Shinder y Cross, 2008](#)).

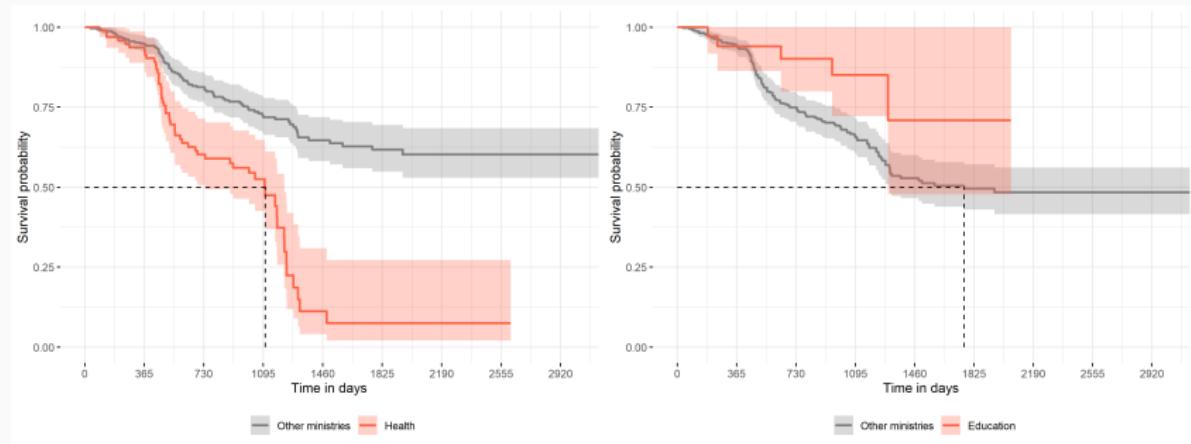
Desempeño y precisión del algoritmo OCR



Validación de la variable sexo



Curvas Kaplan-Meier para altos directivos públicos



Fuente: Elaboración con base en González-Bustamante et al. (2020).

NLP para eventos en América Latina

Revisión automatizada de reportes de prensa

| Lote | Archivos | Años | Formato | Tamaño |
|---------|--|------|--------------|---------|
| Lote 01 | LAWR-1975 a LAWR-1979 Incluyendo 1997 | 5 | CSV UTF-8 | 9,56 MB |
| Lote 02 | LAWR-1980 a LAWR-1998 Excluyendo 1997 Incluyendo primera mitad de 2003 | 18,5 | Imágenes PNG | 38,5 GB |
| Lote 03 | LAWR-1999 a LAWR-2021 Excluyendo primera mitad de 2003 | 22,5 | Archivos PDF | 339 MB |

Nota: Revisión de 12 países de América Latina gracias a St Hilda's College Muriel Wise Fund.

Fuente: Elaboración con base en [González-Bustamante \(2021a\)](#).

Detalles logísticos: Documentación/Metadata

 Repositorio permanente con un **Digital Object Identifier (DOI)**, aunque existe período de embargo (dos años).

Data Management Plan (DMP) registrado con detalles y procedimientos técnicos.

Control de versiones Git conectado con  **Open Science Framework (OSF)** con DOI permanente para asegurar replicabilidad de todo el  código programado cuando se libere, información de los paquetes utilizados y las versiones de  y Python.

Detalles logísticos: Almacenamiento

Repositorios de trabajo alojados de forma privada con control de versiones en  GitHub. Acceso es controlado con autentificación criptográfica con  **two-factor authentication (2FA)** con dispositivos físicos de seguridad USB que emiten **one-time passwords** para generar una autentificación criptográfica basada en protocolos Fast IDentity Online (FIDO) 2 y Universal 2nd Factor (U2F).

Respaldados en  **Hierarchical File Server (HFS)** para su recuperación en casos de incidentes no previstos, también controlado con  2FA. Estos archivos de seguridad se encuentran en un hub encriptado con un cifrado 256-bit AES, por tanto, soporta el almacenamiento de información pública, datos institucionales internos e información confidencial.

Respaldo automático diario, mantenimiento ligero semanal y profundo cada 28 días. Procedimiento programado hasta mayo de 2024.

Detalles logísticos: Accesos y seguridad

El etiquetado de datos para el entrenamiento de algoritmos de machine learning se realiza en una  **plataforma en línea** conectada con los repositorios con información en formato JavaScript Object Notation (JSON).

La plataforma de etiquetado cuenta con **protección de HIPPA** y además cumple con California Consumer Privacy Act (CCPA) y General Data Protection Regulation (GDPR).

El acceso de los codificadores es restringido mediante  contraseñas seguras solo a tareas de etiquetado y al terminar el proceso sus credenciales fueron revocadas inmediatamente.

Detalles logísticos: Almacenamiento de largo plazo

La base de datos final y su documentación serán depositados en el  repositorio digital de **Bodleian Libraries** de University of Oxford.

El código programado será liberado y se incorporará al  **GitHub Archive Program**. Como parte de este programa a comienzos de 2020 instantáneas físicas con códigos públicos fueron archivados en película endurecida (1.000 años de duración) y depositados en  **GitHub Artic Code Vault** en una mina de carbón clausurada en una montaña en Noruega. Este procedimiento se repetirá cada cierto tiempo.

Detalles logísticos: Liberación de datos

El acceso a los archivos de LAWR se realizó a través de distintas suscripciones institucionales como se mencionó previamente, por tanto, estos datos no serán liberados y se almacenan con altos estándares de seguridad.

La información elaborada con algoritmos de OCR y machine learning, los datos y su documentación serán liberados en octubre de 2023 para su reutilización bajo una licencia **Creative Commons Attribution 4.0 International (cc BY 4.0)**.

Esta licencia de acceso abierto permite compartir, reutilizar y readaptar los datos sin restricciones adicionales siempre y cuándo se otorgue el reconocimiento apropiado.

Algoritmos de binarización de imágenes

| POSTSCRIPT | | | | | |
|--|-------|-------|-------|--------|--------|
| POLITICS | | | | | |
| VENEZUELAN CABINET* | | | | | |
| Foreign affairs: Ernesto Leal ¹ Finance: Emilio Pérez Economy and development: Julio Cárdenas Central bank: Silvio de la Torre Agriculture: Roberto Pimentel Transport and communication: Jaime Isidro Labour: Carlos Hernández Justice: Alfredo Duhamel Trade: Carlos Víctor Gamboa ¹ Information: Andrés Eloy Blanco Finance: General Fernando Ochoa Arias ² Health: Ernesto Sotomayor Labour: Francisco Rosales Central bank: Silvio de la Torre Attorney general: Guillermo Vargas Sandoval *Sworn in by President Vicente Chávez **New appointment. †Last cabinet WR-90-16 | | | | | |
| PERUVIAN VIOLENCE | | | | | |
| Death rate | MIL | Ch. | Subs. | Narcos | Total |
| 1987 | 106 | 166 | 165 | 165 | 517 |
| 1988 | 269 | 1,036 | 667 | 1,986 | |
| 1989 | 348 | 1,456 | 1,251 | 149 | 3,198 |
| 1990 | 258 | 1,594 | 1,251 | 68 | 3,452 |
| 1991 | 335 | 1,216 | 1,519 | 36 | 3,290 |
| Total | 7,261 | 5,767 | 5,162 | 257 | 12,533 |
| Monthly average in administration of: Fernando Belaúnde 130 Alan García 156 Alberto Fujimori 252 Source: Defense ministry and senate. | | | | | |

| POSTSCRIPT | | | | | |
|--|-------|-------|-------|--------|--------|
| POLITICS | | | | | |
| VENEZUELAN CABINET* | | | | | |
| Foreign affairs: Ernesto Leal ¹ Finance: Emilio Pérez Economy and development: Julio Cárdenas Central bank: Silvio de la Torre Agriculture: Roberto Pimentel Transport and communication: Jaime Isidro Labour: Carlos Hernández Justice: Alfredo Duhamel Trade: Carlos Víctor Gamboa ¹ Information: Andrés Eloy Blanco Finance: General Fernando Ochoa Arias ² Health: Ernesto Sotomayor Labour: Francisco Rosales Central bank: Silvio de la Torre Attorney general: Guillermo Vargas Sandoval *Sworn in by President Vicente Chávez **New appointment. †Last cabinet WR-90-16 | | | | | |
| PERUVIAN VIOLENCE | | | | | |
| Death rate | MIL | Ch. | Subs. | Narcos | Total |
| 1987 | 106 | 166 | 165 | 165 | 517 |
| 1988 | 269 | 1,036 | 667 | 1,986 | |
| 1989 | 348 | 1,456 | 1,251 | 149 | 3,198 |
| 1990 | 258 | 1,594 | 1,251 | 68 | 3,452 |
| 1991 | 335 | 1,216 | 1,519 | 36 | 3,290 |
| Total | 7,261 | 5,767 | 5,162 | 257 | 12,533 |
| Monthly average in administration of: Fernando Belaúnde 130 Alan García 156 Alberto Fujimori 252 Source: Defense ministry and senate. | | | | | |

| ECONOMY | | | | | |
|-----------------------------------|---------------|-------|-------|-------|--|
| BRAZILIAN OIL Investments | | | | | |
| | US\$bn | | | | |
| 1987 | 1.50 | | | | |
| 1988 | 1.90 | | | | |
| 1989 | 4.40 | | | | |
| 1990 | 5.60 | | | | |
| 1991 | 6.00 | | | | |
| Total | 18.90 | | | | |
| TRADE | | | | | |
| Exports | US\$bn | | | | |
| 1987 | 1.4 | | | | |
| 1988 | 1.9 | | | | |
| 1989 | 2.1 | | | | |
| 1990 | 1.205 | | | | |
| 1991 | 4.15 | | | | |
| Refining Targets | mlpd | | | | |
| 1990 | 450,000 | | | | |
| Exports | 200,000 | | | | |
| 1991 cont. | 110,000 | | | | |
| Source: Energy and mines ministry | Petróbras | | | | |
| COLOMBIAN OIL Exports | | | | | |
| 1987 | 1.2 | | | | |
| 1988 | 1.2 | | | | |
| 1989 | 1.2 | | | | |
| 1990 | 1.2 | | | | |
| 1991 | 1.2 | | | | |
| Refining Targets | mlpd | | | | |
| 1990 | 250,000 | | | | |
| Exports | 200,000 | | | | |
| 1991 cont. | 110,000 | | | | |
| Source: Petróleos | Petróleos | | | | |
| CHILEAN BALANCE | | | | | |
| Exports | US\$bn | | | | |
| 1988 | 7,962 | 8,285 | 8,310 | 8,902 | |
| Imports | 4,833 | 6,502 | 7,037 | 7,419 | |
| Balance | 2,219 | 1,578 | 1,273 | 1,483 | |
| Source: Banco Central | Banco Central | | | | |
| BRAZILIAN OIL Exports | | | | | |
| 1987 | 1.4 | | | | |
| 1988 | 1.9 | | | | |
| 1989 | 2.1 | | | | |
| 1990 | 1.205 | | | | |
| 1991 | 4.15 | | | | |
| Refining Targets | mlpd | | | | |
| 1990 | 450,000 | | | | |
| Exports | 200,000 | | | | |
| 1991 cont. | 110,000 | | | | |
| Source: Energy and mines ministry | Petróleos | | | | |
| COLOMBIAN OIL Exports | | | | | |
| 1987 | 1.2 | | | | |
| 1988 | 1.2 | | | | |
| 1989 | 1.2 | | | | |
| 1990 | 1.2 | | | | |
| 1991 | 1.2 | | | | |
| Refining Targets | mlpd | | | | |
| 1990 | 250,000 | | | | |
| Exports | 200,000 | | | | |
| 1991 cont. | 110,000 | | | | |
| Source: Petróleos | Petróleos | | | | |
| CHILEAN BALANCE | | | | | |
| Exports | US\$bn | | | | |
| 1988 | 7,962 | 8,285 | 8,310 | 8,902 | |
| Imports | 4,833 | 6,502 | 7,037 | 7,419 | |
| Balance | 2,219 | 1,578 | 1,273 | 1,483 | |
| Source: Banco Central | Banco Central | | | | |

LATIN AMERICAN WEEKLY REPORT is published weekly 50 issues a year by Latin American Newsletters, 41 Old Street, London EC1V 9EW, England. Telephone 071 251 0012. Fax 071 253 6193. EDITOR: EDUARDO CRAWLEY. Latin American Newsletters (Journals) Ltd., 40 The Oval, London SE17 3RS, UK. ISSN 0263 6193. © 1992 Latin American Newsletters Ltd. All rights reserved. No part of the publication may be reproduced, stored in any retrieval system or transmitted, in whole or in part, in any form or by any means, electronic, mechanical, photocopying or otherwise, without permission of the publisher. Registered as a newspaper at the Post Office, London. ISSN 0263 6193. The current series of reports is identified by initial letters for the subjects, two hours for the year and a letter for Periodical. This is Week 19 of the Weekly Report, 1992, issue number 4. ISSN 0263 6193. Address reprint requests to Latin American Newsletters Ltd., 40 The Oval, London SE17 3RS, UK. ISSN 0263 6193. © 1992 Latin American Newsletters Ltd. All rights reserved. No part of the publication may be reproduced, stored in any retrieval system or transmitted, in whole or in part, in any form or by any means, electronic, mechanical, photocopying or otherwise, without permission of the publisher. Registered as a newspaper at the Post Office, London. ISSN 0263 6193. The current series of reports is identified by initial letters for the subjects, two hours for the year and a letter for Periodical. This is Week 19 of the Weekly Report, 1992, issue number 4. ISSN 0263 6193.

LATIN AMERICAN WEEKLY REPORT is published weekly 50 issues a year by Latin American Newsletters, 41 Old Street, London EC1V 9EW, England. Telephone 071 251 0012. Fax 071 253 6193. EDITOR: EDUARD CRAWLEY. Latin American Newsletters (Journals) Ltd., 40 The Oval, London SE17 3RS, UK. ISSN 0263 6193. © 1992 Latin American Newsletters Ltd. All rights reserved. No part of the publication may be reproduced, stored in any retrieval system or transmitted, in whole or in part, in any form or by any means, electronic, mechanical, photocopying or otherwise, without permission of the publisher. Registered as a newspaper at the Post Office, London. ISSN 0263 6193. The current series of reports is identified by initial letters for the subjects, two hours for the year and a letter for Periodical. This is Week 19 of the Weekly Report, 1992, issue number 4. ISSN 0263 6193.

LATIN AMERICAN WEEKLY REPORT is published weekly 50 issues a year by Latin American Newsletters, 41 Old Street, London EC1V 9EW, England. Telephone 071 251 0012. Fax 071 253 6193. EDITOR: EDUARD CRAWLEY. Latin American Newsletters (Journals) Ltd., 40 The Oval, London SE17 3RS, UK. ISSN 0263 6193. © 1992 Latin American Newsletters Ltd. All rights reserved. No part of the publication may be reproduced, stored in any retrieval system or transmitted, in whole or in part, in any form or by any means, electronic, mechanical, photocopying or otherwise, without permission of the publisher. Registered as a newspaper at the Post Office, London. ISSN 0263 6193. The current series of reports is identified by initial letters for the subjects, two hours for the year and a letter for Periodical. This is Week 19 of the Weekly Report, 1992, issue number 4. ISSN 0263 6193.

Algoritmos de binarización de imágenes

Nota: Página pre-binariazación y luego del algoritmo de Singh et al. (2011).

Algoritmos de binarización de imágenes

Nota: Página pre-binariación y luego del algoritmo de Su et al. (2010).

Módulo de análisis de diseño de Tesseract

Durante la etapa de layout analysis resulta particularmente desafiante detectar ciertos contornos que de tablas y columnas (Shafait y Smith, 2010; Smith, 2009).

Para manejar esta dificultad utilizamos el layout analysis de Tesseract aplica un proceso de **tab-stop detection** de varios pasos que permite identificar columnas y luego segmentar los párrafos del texto en bloques.

El algoritmo se aplica sobre el segundo y tercer lote de archivos, lo que implica aplicar el algoritmo sobre archivos de 41 años. Tanto imágenes escaneadas en formato PNG como los documentos en PDF se convierten en texto plano en formato Unicode (TXT UTF-8) que luego se convierte a CSV UTF-8.

Módulo de análisis de diseño de Tesseract



How Domingo Cavallo rose to become Menem's virtual prime minister

Argentines are still hotly debating what lay behind the sudden resignation of economy minister Antonio Flores, which was triggered by a speech made by Domingo Cavallo (WR-91-08). The simple explanation—the rapid, uncontrolled upsurge of the inflation rate—is a plausible, decent reason, and much has been published about all sorts of political plots.

One thing seems certain: that González had several times waved his resignation at Menem—and, according to sources close to him, that he knew President Carlos Menem would one day accept it. Indeed, a widely accepted version is that the former president instructed his report Cavallo as his economy minister, to preside over the period of transition in which Menem would take a three-phase process of succession. In this scenario, all González did was to choose his own timing.

Another strand of interpretation is that the change took place as Menem suddenly lost his confidence in Flores, who had not bargained with the depth of adverse popular feeling towards his decision to pardon the leaders of the coup of December 1976. At the time, the scandal over widespread corruption in the ministries of the previous government, and the reaches of the government, and the reach of the economy minister, were so great, that the President's brother, and Carlos Menem's most trusted adviser,

This, at a time when even such highly placed members of his government as Vice-President Eduardo Duhalde, were interested in the manoeuvring to engineer Menem's departure from the presidency for what was speculation at the time, but now appears to be true, that US diplomats, and even US diplomats, have been exploring since late 1990. ■



How Domingo Cavallo rose to become Menem's virtual prime minister

Argentines are still hotly debating what lay behind the sudden resignation of economy minister Antonio Flores, which was triggered by a speech made by Domingo Cavallo (WR-91-08). The simple explanation—the rapid, uncontrolled upsurge of the inflation rate—is a plausible, decent reason, and much has been published about all sorts of political plots.

One thing seems certain: that González had several times waved his resignation at Menem—and, according to sources close to him, that he knew President Carlos Menem would one day accept it. Indeed, a widely accepted version is that the former president instructed his report Cavallo as his economy minister, to preside over the period of transition in which Menem would take a three-phase process of succession. In this scenario, all González did was to choose his own timing.

It was at this point, the scenario goes, that Cavallo bolstered Menem's resolve to hang on. As a consequence, Menem appointed him as the country's new economy minister, but as a virtual prime minister—able to organise the share-out of the ministry's departments. At the time, the scandal over widespread corruption in the ministries of the previous government, and the reaches of the economy minister, were so great, that the President's brother, and Carlos Menem's most trusted adviser,

This was at this point, the scenario goes, that Cavallo bolstered Menem's resolve to hang on. As a virtual prime minister, he was not only as the new economy minister, but as a virtual prime minister—able to organise the share-out of the ministry's departments. At the time, the scandal over widespread corruption in the ministries of the previous government, and the reaches of the economy minister, were so great, that the President's brother, and Carlos Menem's most trusted adviser,

Many expect Cavallo to push for an agreement with the main opposition party, Raúl Alfonsín's Union Cívica Radical, with whom he has contacts. He may even be interested in some form of coalition that would ensure the continuity of the government. This is an option that many politicians, economists, business leaders, and even US diplomats, have been exploring since late 1990. ■

PERU: Wildfire spread of cholera epidemic. (2)

EQUADOR: Holiday 'holiday' will impact on Miraflores. (2)

COLOMBIA: Prospects of talks, abroad, with CND. (2)

VENEZUELA: Lower prices force budget re-think. (3)

BOLIVIA: A Soviet partner for oil exploration. (2)

PARAGUAY: Placing limits on democracy. (4)

CHILE: Proposal to extend Aylwin's mandate. (8)

REPUBLIC: Watch out for the Gulf affair. (7)

HAITI: Aristide faces first major challenges. (11)

EL SALVADOR: Setback as FMLN 'breaches security'. (10/2/11)

PANAMA: Time to recall Noriega officers? (11)

Nota: Aplicación de líneas tab-stop y estructura de columnas.

Módulo de análisis de diseño de Tesseract

The image shows a newspaper clipping from the "Latin American weekly report". The masthead includes the text "Latin American", "weekly report", and "21 February 1991, Vol.91-07". The main headline reads "How Domingo Cavallo rose to become Menem's virtual prime minister". The text discusses the political maneuvering behind Cavallo's appointment, mentioning Menem's loss of confidence in his economy minister and the subsequent widespread corruption. It also notes the replacement of Eduardo Menem by Cavallo. The bottom of the page lists various international news items such as Peru, Venezuela, Brazil, Bolivia, Chile, El Salvador, Costa Rica, and Haiti.

Year and Week

Headline

Paragraphs

Nota: Identificación y segmentación de bloques.

Conjunto de datos dependiente del tiempo

| Ministro | Entrada | Salida | Evento | Cuest. | Re-Elección | Legislativo |
|----------|---------|--------|--------|--------|-------------|-------------|
| 1 | 0 | 30 | 0 | 0 | 1 | 3 |
| 1 | 30 | 60 | 0 | 0 | 1 | 3 |
| 1 | 60 | 80 | 0 | 1 | 1 | 4 |
| 2 | 0 | 15 | 0 | 0 | 1 | 3 |
| 2 | 15 | 30 | 1 | 1 | 1 | 3 |
| 2 | 30 | 60 | 0 | 1 | 1 | 3 |
| 2 | 60 | 80 | 2 | 2 | 1 | 4 |

Fuente: Elaboración con base en [González-Bustamante \(2020, 2021a\)](#).

Codificación de datos dependientes del tiempo

Por ejemplo, en [González-Bustamante \(2021b\)](#) fusionamos los conjuntos de datos de ministros de Brasil y Chile entre 1990 y 2014 de [Franz y Codato \(2016\)](#) and [González-Bustamante y Olivares \(2018\)](#).

De esta forma obtenemos un conjunto de **488 observaciones** que codificamos como **dependiente del tiempo** con puntos de corte trimestrales para todo el período con el fin de incorporar la aprobación presidencial y datos macroeconómicos como **time-varying covariates**.

La base se codifica con casos que tienen múltiples observaciones, en este caso i -th ministros, según intervalos de tiempo definidos correspondientes a eventos específicos: los cuatro trimestres de cada año. La varianza de las covariables que varían en el tiempo se codifica sobre el intervalo cerrado.

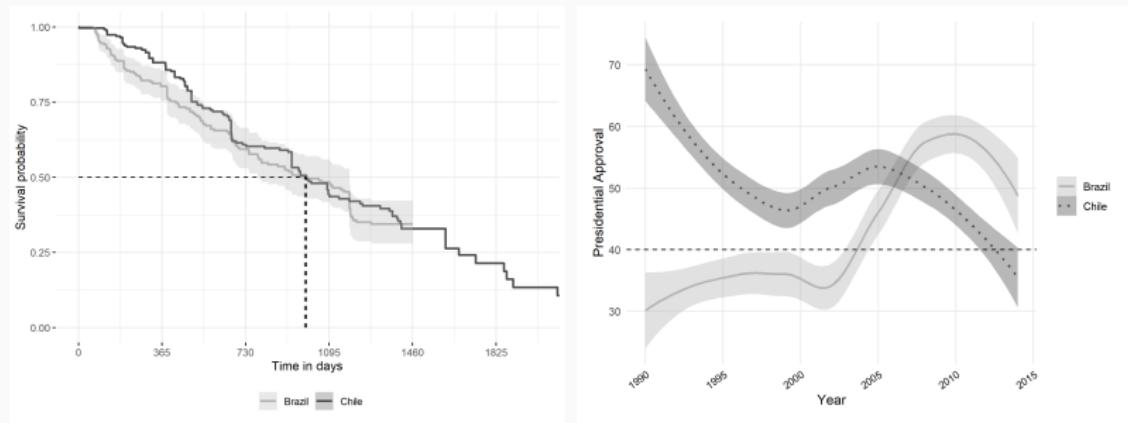
Codificación de datos dependientes del tiempo

El último intervalo de cada caso termina con la salida (u otro *competing risk*) del ministro Y_i , excepto en los casos censurados administrativamente. Fusionamos en cada intervalo $Z(t)$ **quarterly smoothed presidential approval** con datos de [Carlin et al. \(2019\)](#).

Además, agregamos indicadores macroeconómicos de [World Bank \(2020\)](#) (crecimiento e inflación) y el número efectivo de partidos con el indicador de [Gallagher y Mitchell \(2005\)](#) actualizado.

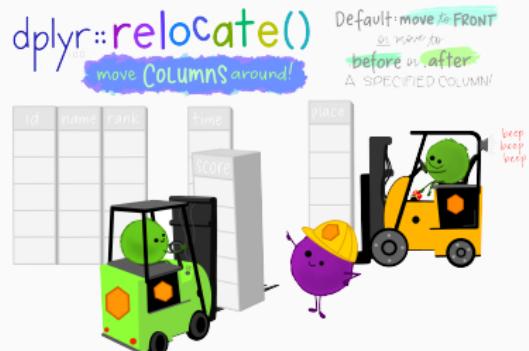
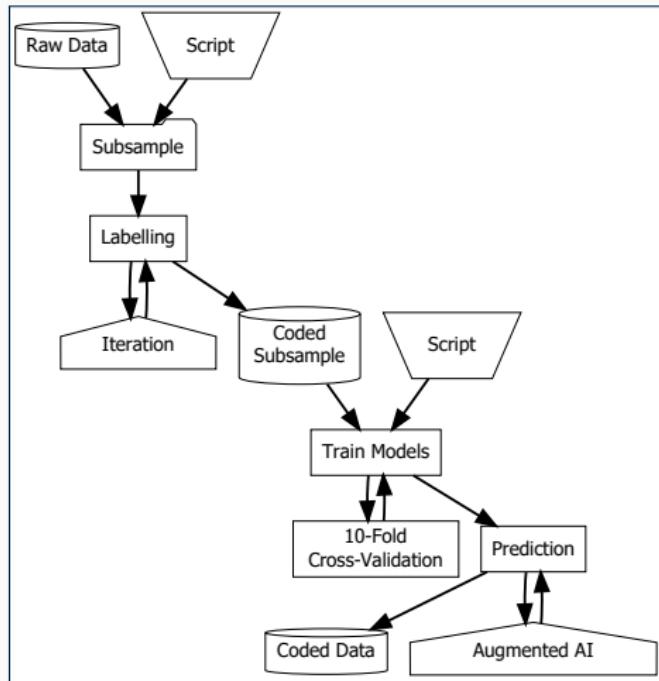
Esto nos permite obtener un conjunto con **4.245 observaciones**.

Estimaciones de supervivencia y aprobación



Fuente: Elaboración con base en [González-Bustamante \(2021b\)](#).

NLP en LAWR y modelos supervisados para identificar crisis



Indicadores del proceso de etiquetado

| Codificador | Etiquetas | $M(t)$ etiqueta | \sum Tiempo | Consenso |
|-------------|-----------|-----------------|---------------|----------|
| 1 | 966 | 39s | 10h 25m | 94.05 |
| 2 | 718 | 1m 15s | 15h 0m | 94.84 |
| 3 | 704 | 38s | 7h 29m | 94.82 |
| 4 | 318 | 1m 4s | 5h 38m | 95.47 |
| 5 | 263 | 17s | 1h 14m | 99.26 |
| 6 | 101 | 1m 17s | 2h 9m | 88.50 |

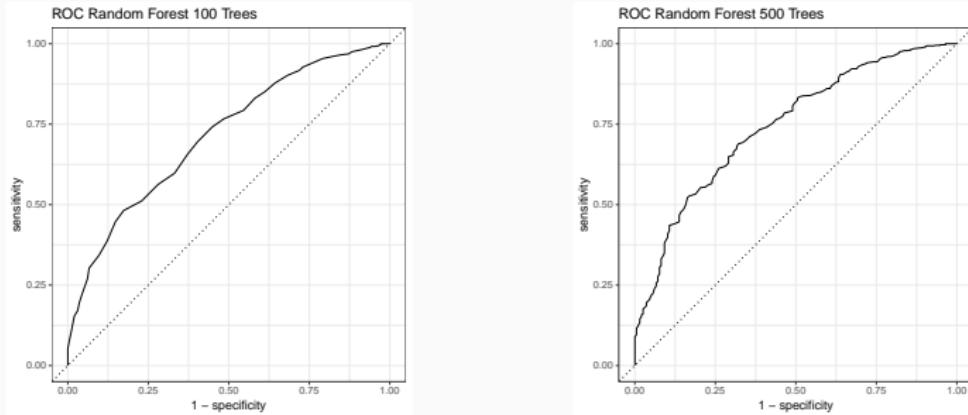
Fuente: Elaboración con base en [González-Bustamante \(2021a\)](#).

Desempeño de los modelos supervisados

| Modelo | Exactitud | Precisión | Sensibilidad |
|-----------------------------|-----------|-----------|--------------|
| Naïve Bayes | 0,714 | 0,803 | 0,825 |
| SVM Kernel-Lineal | 0,658 | 0,802 | 0,728 |
| SVM Kernel-Gaussiano | 0,756 | 0,756 | 0,999 |
| Random Forest (100 árboles) | 0,761 | 0,760 | 0,998 |
| Random Forest (500 árboles) | 0,761 | 0,760 | 0,999 |

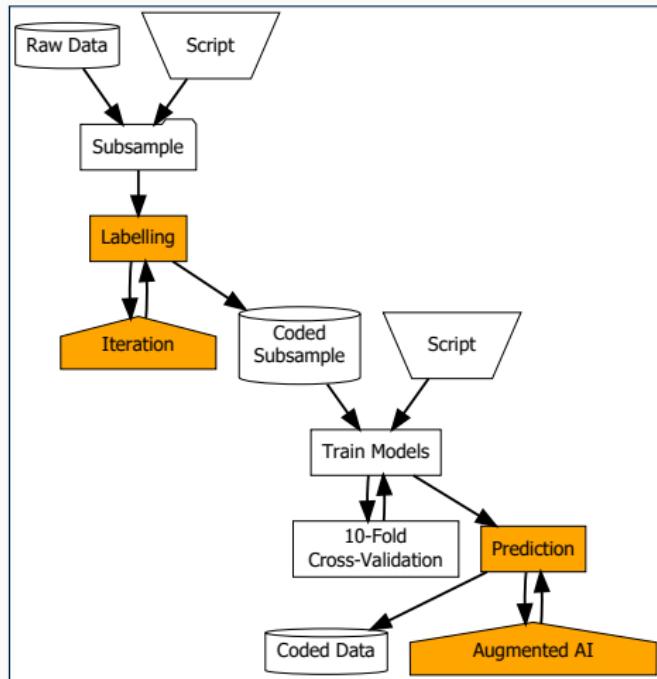
Fuente: Elaboración con base en [González-Bustamante \(2021a\)](#).

Curvas ROC de modelos Random Forest



Fuente: Elaboración con base en [González-Bustamante \(2021a\)](#).

¿Cómo aumentar la capacidad predictiva?



Training Data Lab

Training Data Lab

© 2020 **Training Data Lab** es un grupo de investigación que se enfoca en aplicaciones de ciencia de datos en ciencias sociales en tres áreas interconectadas: **minería de datos, modelamiento econométrico y aprendizaje automático**. Por una parte, buscamos recoger datos con técnicas de minería para elaborar modelos econométricos con técnicas observacionales o de emparejamiento.

Por otro lado, nos enfocamos en entrenar modelos con técnicas de aprendizaje automático y profundo etiquetando conjuntos de datos para diferentes proyectos. Lo anterior, nos permite clasificar datos no codificados usando nuestros modelos entrenados incorporando validación humana en el flujo de trabajo, lo que mejora la inteligencia artificial en los procesos de aprendizaje.



Universiteit
Leiden



UNIVERSIDAD
MAYOR

Algoritmo clasificador para mociones legislativas

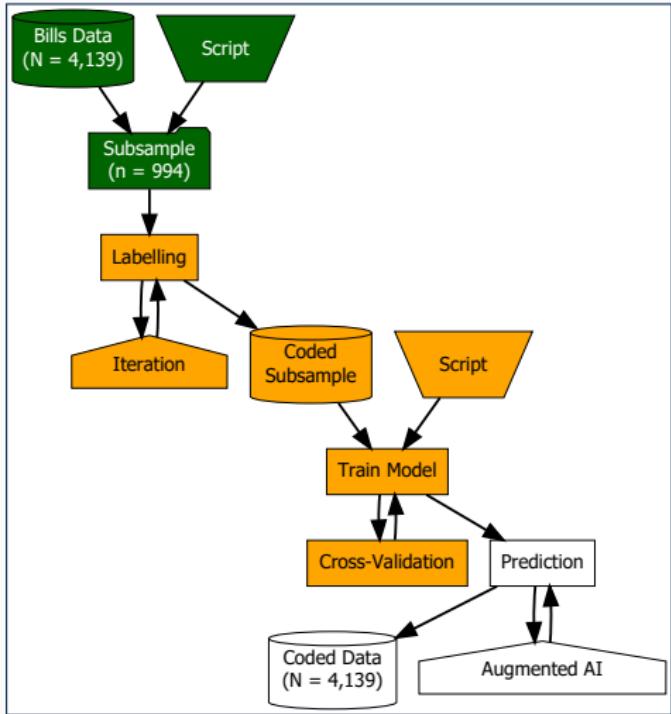
 Bastián González-Bustamante, Carla Cisternas, Diego Aguilar
y Jaquelin Morillo

 training-datalab.com/projects/chilean-congress-bills

Training Data on Chilean Congress Bills. A partir de un conjunto de datos de proyectos de ley de la Cámara de Diputados de Chile entre 2006 y 2018 ($N = 4.139$), período que corresponde a tres administraciones, extraemos una submuestra aleatoria considerando algunos proyectos de ley por mes.

En esta submuestra realizamos dos procedimientos de codificación de datos para identificar tanto el tema del proyecto de ley como su alcance territorial.

Algoritmo clasificador para mociones legislativas



Proyecto VIP de machine learning



APLICACIONES DE MACHINE LEARNING EN POLÍTICAS PÚBLICAS Y ECONOMÍA

Proyecto VIP de machine learning

-  Carla Cisternas y Francisco Castañeda (responsables)
-  Bastián González-Bustamante, Rodrigo Cuevas y Diego Aguilar
-  training-datalab.com/projects/vip-project-machine-learning

Aplicaciones de Machine Learning en políticas públicas y economía.

Este proyecto VIP busca conformar un equipo multidisciplinario de investigación verticalmente integrado para realizar diferentes aplicaciones de aprendizaje automático en temas de políticas públicas y economía. En esta primera versión, nos centraremos en las dinámicas de desinformación en temas económicos y de interés público.

Posteriormente, en futuras versiones, esperamos abordar y analizar otros fenómenos relacionados con la formulación de políticas públicas sectoriales y regulación de mercados.

Proyecto VIP de machine learning

twConstitution



Twitter Online Tracker of the Chilean Referendum for a New Constitution

[View the Project on GitHub](#)
bgonzalezbustamante/twConstitution

Twitter Online Tracker of the Chilean Referendum for a New Constitution

version v1.2.6 issues 1 open issues 4 closed DOI 10.17605/OSF.IO/73NDB

license CC-BY-4.0 made with R v4.0.2 made with Jekyll

This is a Twitter online tracker of the Chilean referendum for a new Constitution in October 2020, which contains daily datasets on **#Apruebo** (see words network, *forthcoming*) and **#Rechazo** (see words network, *forthcoming*) viewpoints on this social media.

Data sets are scraped and uploaded regularly. Some of the variables are date, hour, username, tweet text, RT count, fav count, location, among others. The data was collected during the afternoon each day. A couple of exceptions, such as October 6th and 26th, were collected early morning on the following day. This is not an issue because the data could be sliced, and in order to work with the whole period, it is necessary to merge the sets and retain unique cases.

| Dataset | Date | Year | N | Size | Format |
|----------|---------|------|--------|---------|---------------------|
| #Apruebo | Nov. 01 | 2020 | 45,195 | 46.5 MB | CSV |
| #Rechazo | Nov. 01 | 2020 | 22,142 | 24.0 MB | CSV |
| #Apruebo | Oct. 31 | 2020 | 50,244 | 52.1 MB | CSV |
| #Rechazo | Oct. 31 | 2020 | 27,331 | 29.9 MB | CSV |
| #Apruebo | Oct. 30 | 2020 | 53,889 | 56.1 MB | CSV |
| #Rechazo | Oct. 30 | 2020 | 33,479 | 36.6 MB | CSV |

Agradecimientos y referencias

Agradecimientos I

Senior Public Managers: A Novel Dataset on Members of the Chilean Civil Service

Agradezco a Matías Astete y Berenice Orvenes, coautores del artículo que además se hicieron cargo principalmente de la validación de información y recolección de fuentes para la aplicación de algoritmos.

Esta investigación fue financiada por la Facultad de Administración y Economía de la Universidad de Santiago de Chile.

Agradecimientos II

Stability during Presidential Approval Crises in Brazil and Chile

Agradezco a Petra Schleiter, Stephen Whitefield y Moshe Ben Hamo Yeger sus importantes comentarios teóricos y metodológicos. También agradezco a Adriano Codato, Renato Perissinotto, Carla Cisternas y Cristóbal Rovira sus valiosos comentarios versiones preliminares del diseño.

Esta investigación fue financiada por la Agencia Nacional de Investigación y Desarrollo de Chile (ANID/PFCHA/72200340).

Agradecimientos III

A Political Events Data Set of 12 Latin American Presidential Democracies

Mis agradecimientos a Petra Schleiter por sus sustanciales comentarios y apoyo. También agradezco a Carla Cisternas y Diego Aguilar por su ayuda durante el proceso de revisión y validación de archivos. Finalmente, también mis agradecimientos para quienes apoyaron el proceso de etiquetado, especialmente para Jaquelin Morillo.

Esta investigación es financiada por la Agencia Nacional de Investigación y Desarrollo de Chile (ANID/PFCHA/72200340) y St Hilda's College Muriel Wise Fund.

Referencias I

- Carlin, R. E., Hartlyn, J., Hellwig, T., Love, G. J., Martínez-Gallardo, C., y Singer, M. M. (2019). Executive Approval Database 2.0. Dataset, Executive Approval Project. Disponible en <http://www.executiveapproval.org/>.
- Franz, P. y Codato, A. (2016). Ministros de Estado: governos FHC e Lula (1995-2010). Dataset, Observatório de Elites Políticas e Sociais do Brasil, Universidade Federal do Paraná. Disponible en <http://observatory-elites.org/>.
- Gallagher, M. y Mitchell, P. (2005). *The Politics of Electoral Systems*. Oxford University Press, Nueva York.
- González-Bustamante, B. (2020). *Critical Events and Ministerial Turnover in Latin American Presidential Democracies. Application for Transfer of Status*. PhD thesis, Department of Politics and International Relations, St Hilda's College, University of Oxford, Oxford.
- González-Bustamante, B. (2021a). A Political Events Data Set of 12 Latin American Presidential Democracies. Working Paper, Department of Politics and International Relations, St Hilda's College, University of Oxford.
- González-Bustamante, B. (2021b). Ministerial Stability During Presidential Approval Crises: The Moderating Effect of Ministers' Attributes on Dismissals in Brazil and Chile. En *XXVI World Congress of Political Science*, Lisboa.

Referencias II

- González-Bustamante, B., Astete, M., y Orvenes, B. (2020). Altos directivos públicos: Un nuevo conjunto de datos de miembros del servicio civil chileno. *Revista de Gestión Pública*, IX(2):151–169.
- González-Bustamante, B., Astete, M., y Orvenes, B. (2021). Senior Public Managers: A Novel Dataset on Members of the Chilean Civil Service. SocArXiv DOI: 10.31235/osf.io/vshcz.
- González-Bustamante, B. y Olivares, A. (2018). La élite política gubernamental en Chile: Supervivencia de ministros. En Codato, A. y Espinoza, F., editores, *Las élites en las Américas: Diferentes perspectivas*. Editora Universidade Federal do Paraná, Curitiba.
- Radford, B. J. (2019). Automated dictionary generation for political event coding. *Political Science Research and Methods*, 9(1):157–171.
- Sauvola, J. y Pietikäinen, M. (2000). Adaptive document image binarization. *Pattern Recognition*, 33(2):225–236.
- Shafait, F. y Smith, R. (2010). Table detection in heterogeneous documents. En *DAS '10: Proceedings of the 9th IAPR International Workshop on Document Analysis Systems*, Boston. ACM Press.

Referencias III

- Shinder, L. y Cross, M. (2008). Understanding Cybercrime Prevention. En Cross, M., editor, *Scene of the Cybercrime*. Elsevier, Burlington.
- Singh, T. R., Roy, S., Singh, O. I., Sinam, T., y Singh, K. M. (2011). A New Local Adaptive Thresholding Technique in Binarization. *IJCSI International Journal of Computer Science Issues*, 8(6-2):271–277.
- Smith, R. (2007). An Overview of the Tesseract OCR Engine. En *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007) Vol 2*. IEEE.
- Smith, R. W. (2009). Hybrid Page Layout Analysis via Tab-Stop Detection. En *2009 10th International Conference on Document Analysis and Recognition*, Barcelona. IEEE.
- Su, B., Lu, S., y Tan, C. L. (2010). Binarization of historical document images using the local maximum and minimum. En *DAS '10: Proceedings of the 9th IAPR International Workshop on Document Analysis Systems*, Boston. ACM Press.
- Wais, K. (2016). Gender Prediction Methods Based on First Names with genderizeR. *The R Journal*, 8(1):17–37.
- World Bank (2020). World Bank Open Data. Dataset, World Bank. Disponible en <https://data.worldbank.org/>.

Contacto

Bastián González-Bustamante

DPhil (PhD) Researcher

Department of Politics and International Relations

& St Hilda's College

University of Oxford

📍 St Hilda's College, Cowley Place, Oxford OX4 1DY

✉️ bastian.gonzalezbustamante@politics.ox.ac.uk

🏠 <https://bgonzalezbustamante.com>

Profesor Instructor

Departamento de Gestión y Políticas Públicas

Facultad de Administración y Economía

Universidad de Santiago de Chile

📍 Av. Lib. B. O'Higgins 3363, Estación Central, Santiago

✉️ bastian.gonzalez.b@usach.cl



Presentación compilada con **LATEX** y algunos ☕

⌚ Descargar la versión más reciente desde [GitHub](#)

♾ Artwork utilizado disponible en [GitHub](#)

Muchas gracias por su atención

